MINIMAX WAVELET ESTIMATIONS VIA BLOCK THRESHOLDING

by

T. Tony Cai
Purdue University

Technical Report #96-41

Department of Statistics
Purdue University
West Lafayette, IN USA

November 1996

*

Minimax Wavelet Estimation Via Block Thresholding

T. Tony Cai
Department of Statistics, Purdue University

October, 1996

Abstract

Wavelet shrinkage methods have been very successful in nonparametric regression. The most commonly used wavelet procedures achieve adaptivity through term-by-term thresholding. The resulting estimators attain the minimax rates of convergence up to a logarithmic factor.

In the present paper, we propose a block thresholding method where wavelet coefficients are thresholded in blocks, rather than individually. We show that the estimators produced by the procedure, *BlockShrink*, are spatially adaptive and asymptotically optimal both for global and local estimation. The *BlockShrink* estimators achieve the exact optimal rates of convergence for global estimation over a range of function classes of inhomogeneous smoothness. The estimators attain the adaptive minimax rates for estimating regression functions at a point. Moreover, a large simulation study shows that the *BlockShrink* estimators yield uniformly better results in terms of the mean squared error than the widely used VisuShrink estimators. The procedure is easy to implement and the computational cost is of order $O(n)$.

Keywords: Minimax Estimation; Nonparametric Regression; Orthogonal Wavelet Bases of Compact Support; Block Thresholding; Spatial Adaptivity.

AMS 1991 Subject Classification: Primary 62G07, Secondary 62G20.

1 Introduction

Consider the nonparametric regression model:

$$y_i = f(x_i) + \epsilon z_i \tag{1}$$

i = 1, 2, ..., n (= 2^J), $x_i = \frac{i}{n}$ and z_i's are iid N(0, 1).

The function f($\cdot$) is an unknown function of interest. In the present paper, we are interested in estimating function f($\cdot$) globally as well as estimating f($\cdot$) locally. We measure the quality of recovery by the mean squared error.

The theory and methods of nonparametric regression have been developed rapidly over the last few years with the introduction of nonlinear wavelet methods by Donoho and Johnstone, et al (see [7] and [10],[9]). Wavelet procedures have demonstrated unprecedented successes in terms of asymptotical optimality, spatial adaptivity and computational efficiency. Wavelet methods achieve their unusual adaptivity through shrinkage of the empirical wavelet coefficients. They enjoy excellent mean squared error properties when used to estimate functions that are only piecewise smooth and have near optimal convergence rates over large function classes. In contrast, traditional linear estimators typically achieve good performance only for relatively smooth functions.

Standard wavelet thresholding procedures are based on term-by-term decisions about wavelet expansions. There, wavelet coefficients are estimated individually, retained or discarded depending on the magnitude of the corresponding empirical wavelet coefficients. The most widely used wavelet shrinkage method in nonparametric regression is the Donoho-Johnstone's VisuShrink procedure ([7], [10]) . The VisuShrink procedure has three steps:

1. Transform the noisy data via the discrete wavelet transform;
2. Denoise the empirical wavelet coefficients by "hard" or "soft" thresholding rules with threshold $\lambda = \epsilon\sqrt{2\log n}$.
3. Estimate function f at the sample points by inverse discrete wavelet transform of the denoised wavelet coefficients.

This procedure is adaptive and easy to implement. And with high probability, VisuShrink estimators are at least as smooth as the target function. The estimators produced by the procedure achieve minimax convergence rates up to a logarithmic penalty over a wide range of function classes.

Despite its considerable advantages, however, it has drawbacks. VisuShrink achieves a degree of tradeoff between variance and bias contributions to the mean squared error.

However, the tradeoff is not optimal. It favors reducing variance over bias. The squared bias is of higher order of magnitude than the variance. In other words, the estimator is over-smoothed. As a result, it creates a logarithmic penalty in the mean squared error. The problem can not be solved by only fine tuning the threshold level. In fact, the threshold rule is asymptotically optimal among all such uniform term-by-term threshold rules.

The difficulty of term-by-term thresholding is caused by the relative inaccuracy that individual wavelet coefficients are estimated. Because of the need to guard against "false positives" about the presence of irregularities of the underlying function *f(x)*, VisuShrink inevitably removes too many terms in the wavelet series. As a result the estimators are too biased and so do not react to relatively subtle changes in the underlying regression function *f(x)*.

To solve the problem, we attempt to use the idea of block thresholding where wavelet coefficients are threshloded in blocks, rather than individually. The idea of block thresholding can be traced back to Efroimovich ([11]) in orthogonal series estimators; and Kerkyacharian, Picard and Tribouley ([15]), for wavelet density estimation. But these block thresholding are not local, so they do not enjoy a high degree of spatial adaptivity. Local versions of block thresholding first appeared in Hall, Kerkyacharian and Picard ([12], [13]). They proposed a block threshold rule and showed that the estimators attain the minimax convergence rates over a range of function classes. However, their simulation study showed that the estimators have little advantage over the VisuShrink estimators when the signal-to-noise ratio is high (see [14]).

In the present paper, we introduce a new block thresholding procedure, *BlockShrink*, which is very easy to implement and has broader spatial adaptivity than the term-by-term thresholding methods. The *BlockShrink* procedure differs from the methods discussed in Hall, Kerkyacharian and Picard ([12], [13]) in the choice of the block length, the estimator of the energy, and the threshold level.

The *BlockShrink* introduced in Section 3 has the following ingredients:

1. Transform the noisy data via the discrete wavelet transform: $\tilde{\Theta} = W \cdot Y$.

2. At each resolution level, group the noisy wavelet coefficients into blocks of length $L \approx \log n$. A block (jb) is deemed to contain significant information about the function f if the energy in the block $\tilde{B}_{(jb)} \equiv \sum_{k \in (jb)} \tilde{\theta}^2_{jk} > 5L\epsilon^2$ and then all the coefficients in the block are retained; otherwise the block is deemed insignificant and all the coefficients in the block are discarded.

3. Obtain the estimate of function *f(x)* at the sample points by the inverse discrete wavelet transform of the denoised wavelet coefficients.

We show in Section 5 that the *BlockShrink* procedure enjoys a high degree of adaptivity and spatial adaptivity. The *BlockShrink* estimators achieve true optimality in terms of convergence rates over a wide range of function classes of inhomogeneous smoothness.

For global estimation, it attains the exact optimal rates of covergence over an interval of function classes $\mathcal{F}$ defined in Section 4; for estimating functions at a point, it achieves the adaptive minimax rates over a broad range of the Hölder classes.

A simulation study (see [4]) showed that the *BlockShrink* estimators uniformly outperform the VisuShrink estimators in terms of the mean squared error, even when the signal-to-noise ratio is high in which case the VisuShrink is known to perform very well. Furthermore, in the cases of high signal-to-noise ratio, the *BlockShrink* also yields uniformly better results than the RiskShrink (see [7]) and the SureShrink (see [9]) in all examples. In some cases, the improvement is substantial. For instance, from Table 1, the *BlockShrink* estimator of Doppler, a function with significant spatial inhomogeneity, achieves better performance with samples of size n than each of the other three estimators with samples of size $2 \cdot n$.

The paper is organized as follows. Section 2 introduces wavelets and some basic notations. The *BlockShrink* procedure is presented in Section 3. Section 4 gives the definitions and properties of the function classes of interest. The optimalities of the procedure are discussed in Section 5. Section 6 discusses the simulation results. All the proofs are postponed to Section 7.

2 Wavelets

Wavelets are a relatively new concept in applied mathematics ([5], [6]). But they have already let to exciting applications in many fields, from signal and image processing to statistical estimation.

Wavelets are a special type of orthonormal basis in L_2 space. An orthonormal wavelet basis is generated from dilations and translations of two basic functions, a "father" wavelet ϕ and a "mother" wavelet ψ. The functions ϕ and ψ are assumed to be compactly supported. also we assume that ψ has D vanishing moments and ϕ satisfies

$$\int \phi = 1$$

Daubechies ([6]) constructed compactly supported wavelets called Coiflets which can have arbitrary number of vanishing moments for both the father wavelet ϕ and mother wavelet ψ. Denote W(D) the collection of Coiflets $\{\phi, \psi\}$ of order D. So if $\{\phi, \psi\} \in W(D)$, then ϕ and ψ are compactly supported and satisfy

$$\int x^i \phi(x) dx = 0, \qquad \text{for i = 1, 2, ..., D-1}$$
$$\int x^i \psi(x) dx = 0, \qquad \text{for i = 0, 1, ..., D-1}$$

Let

$$\phi_{jk}(t) = 2^{j/2}\phi(2^j t - k), \quad \psi_{jk}(t) = 2^{j/2}\psi(2^j t - k)$$

And denote the periodized wavelets

$$\phi_{jk}^p(t) = \sum_{l\in\mathcal{Z}} \phi_{jk}(t-l), \quad \psi_{jk}^p(t) = \sum_{l\in\mathcal{Z}} \phi_{jk}(t-l) \quad \text{for } t \in [0,1]$$

For the purposes of this paper, we use the periodized wavelet bases on [0, 1]. The collection $\{\phi_{j_0k}^p, k = 1,...,2^{j_0}; \psi_{jk}^p, j \geq j_0, k = 1,...,2^j\}$ constitutes such an orthonormal basis of $L_2[0,1]$. Note that the basis functions are periodized at the boundary. The superscript "p" will be suppressed from the notations for convenience. This basis has an associated exact orthogonal Discrete Wavelet Trasnform (DWT) that transforms data into wavelet coefficient domains.

For a given square-integrable function f on [0, 1], denote

$$\xi_{jk} = \langle f, \phi_{jk}\rangle, \qquad \theta_{jk} = \langle f, \psi_{jk}\rangle$$

So the function f can be expanded into a wavelet series:

$$f(x) = \sum_{k=1}^{2^{j_0}} \xi_{j_0k}\phi_{j_0k}(x) + \sum_{j=j_0}^{\infty}\sum_{k=1}^{2^j} \theta_{jk}\psi_{jk}(x) \tag{2}$$

Wavelet transform decomposes a function into different resolution components. In (2), ξ_{j_0k} are the coefficients at the coarsest level. They represent the gross structure of the function f. And θ_{jk} are the wavelet coefficients. They represent finer and finer structures of the function f as the resolution level j increases.

We note that the DWT is an orthogonal transform, so it transforms i.i.d. Gaussian noise to i.i.d. Gaussian noise and it is norm-preserving. This important property of DWT allows us to transform the problem in the function domain into a problem in the sequence domain of the wavelet coefficients with isometry of risks.

3 The BlockShrink Procedure

Given a sample $Y = \{y_i\}$ as in (1). Let $\tilde{\Theta} = W \cdot n^{-1/2}Y$ be the discrete wavelet transform of $n^{-1/2}Y$. Write

$$\tilde{\Theta} = (\tilde{\xi}_{j_01}, \cdots, \tilde{\xi}_{j_02^{j_0}}, \tilde{\theta}_{j_01}, \cdots, \tilde{\theta}_{j_02^{j_0}}, \cdots, \tilde{\theta}_{J-1,1}, \cdots, \tilde{\theta}_{J-1,2^{J-1}})^T$$

Here $\tilde{\xi}_{j_0k}$ are the gross structure terms at the lowest resolution level, and $\tilde{\theta}_{jk}$ ($j = 1, \cdots, J-1, k = 1, \cdots, 2^j$) are fine structure wavelet terms. One may write

$$\tilde{\theta}_{jk} = \theta_{jk} + a_{jk} + n^{-1/2}\epsilon z_{jk} \tag{3}$$

where θ_{jk} is the true wavelet coefficients of f, a_{jk} is some approximation error which is considered "small" by the results of Lemma 1 in Section 4, and z_{jk}'s are the transform of the z_i's and so are i.i.d. $N(0,1)$.

The term-by-term thresholding procedure, VisuShrink, estimates the function f by

$$\hat{f}(x) = \sum_{k=1}^{2^{j_0}} \tilde{\xi}_{j_0 k}\phi_{j_0 k}(x) + \sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j} \tilde{\theta}_{jk} I(|\tilde{\theta}_{jk}| > \epsilon\sqrt{2n^{-1}\log n})\psi_{jk}(x)$$

Here, each wavelet coefficient is estimated separately.

In *BlockShrink*, we threshold wavelet coefficients in groups instead of thresholding individually. At each resolution level j, the empirical wavelet coefficients $\tilde{\theta}_{jk}$ are divided into nonoverlapping blocks of length $L = [\log n]$. Denote (jb) the indices of the coefficients in the bth block at level j, i.e.

$$(jb) = \{(j,k) : (b-1)L + 1 \le k \le bL\}$$

Let $\tilde{B}_{(jb)} = \sum_{(jb)} \tilde{\theta}^2_{jk}$ denote the L_2 energy of the noisy signal in block (jb). From (3), it is clear that $\tilde{B}_{(jb)}$ is biased as an estimate of $\sum_{(jb)} \theta^2_{jk}$, the energy of the function f in the block. In fact,

$$E\tilde{B}_{(jb)} = \sum_{(jb)} (\theta_{jk} + a_{jk})^2 + Ln^{-1}\epsilon^2$$

On the other hand, $\tilde{B}_{(jb)}$ does provide a guideline about the magnitude of $\sum_{(jb)} \theta^2_{jk}$. In *BlockShrink*, a block (jb) is deemed to contain significant information about the function f if $\tilde{B}_{jb} > 5Ln^{-1}\epsilon^2$ and then all the coefficients in the block are retained; otherwise all the coefficients in the block are discarded. Therefore, the coefficients θ_{jk} in the block (jb) is estimated by

$$\hat{\theta}_{jk} = \tilde{\theta}_{jk} I(\tilde{B}_{jb} > 5Ln^{-1}\epsilon^2)$$

And the whole function f is estimated by

$$\hat{f}^*_n(x) = \sum_{k=1}^{2^{j_0}} \tilde{\xi}_{j_0 k}\phi_{j_0 k}(x) + \sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j} \hat{\theta}_{jk}\psi_{jk}(x)$$

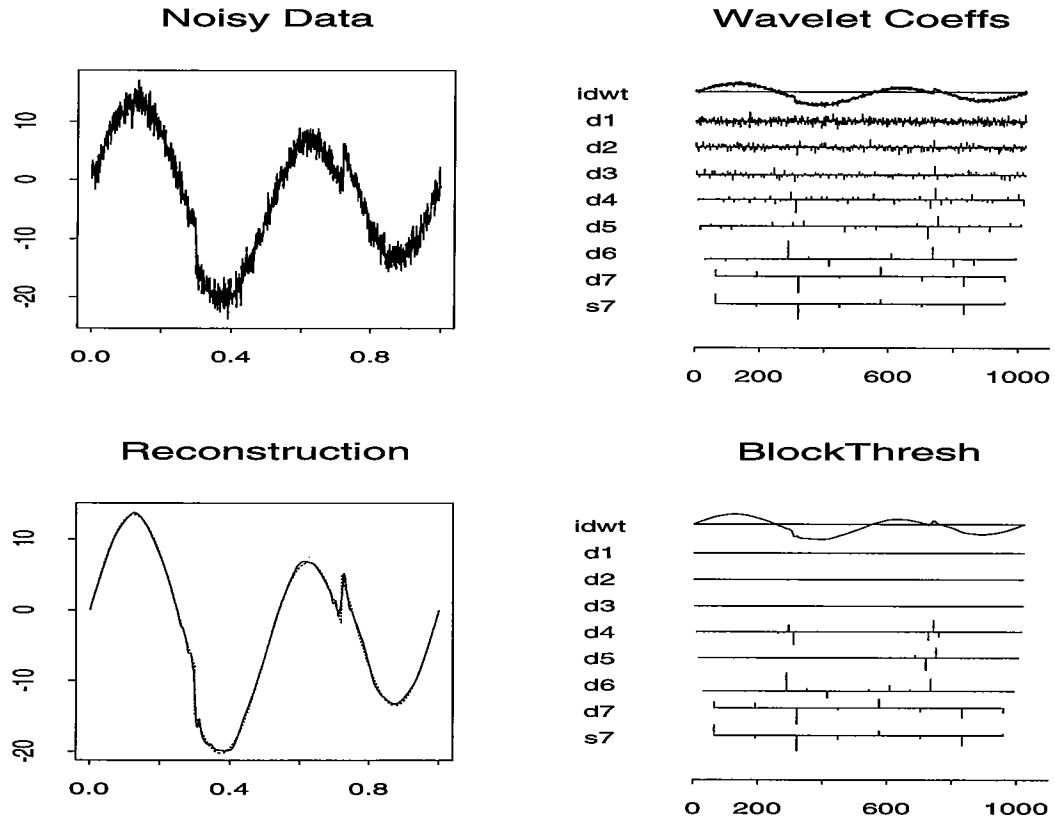
If one is interested in estimating f at the sample points, then the fast Inverse Discrete Wavelet Transform (IDWT) can be used. And $\{f(x_i) : i = 1, \cdots, n\}$ is estimated by $\hat{f} = \{\widehat{f(x_i)} : i = 1, \cdots, n\}$ with

$$\hat{f} = W' \cdot n^{1/2}\hat{\Theta}$$

We call the procedure *BlockShrink*.

We show in Section 5 that the *BlockShrink* estimators attain the exact optimal rates over a wide range of function classes. The *BlockShrink* estimators are appealing not only quantitatively but also qualitatively. They automatically adapt to the smoothness of the target functions. Here is an example of the *BlockShrink* in action. For better comparison,

the true function, HeaviSine, is superimposed on the estimator as dotted line. It is clear that the estimator captures both the smooth and the jump features of the function very well. The reconstruction jumps where the target function jumps; the reconstruction is smooth where the target function is smooth. Further simulation results are discussed in Section 6.



4 The Functions Classes $\mathcal{F}$

Following Hall, Kerkyacharian and Picard's lead, we consider the adaptivity of our *Block-Shrink* procesure over the following function classes which they introduced in [13]. These are not traditional smoothness classes. The function classes contain functions of inhomogeneous smoothness. The functions can be regarded as the superposition of smooth functions with irregular perturbations. See remarks 1 and 2 below. Here is the formal definition of the function classes $\mathcal{F}$.

Definition 1 *Let*

$$\mathcal{F} = \mathcal{F}(\alpha_1, \alpha, \gamma, M_1, M_2, M_3, D, \nu)$$

where $0 \leq \alpha_1 < \alpha \leq D$, $0 \leq \gamma < \frac{1+2\alpha_1}{1+2\alpha}$, *and* $M_1, M_2, M_3, \nu \geq 0$, *denote the class of functions* f *such that for any* $j \geq j_0 > 0$ *there exists a set of integers* A_j *for which all of the following are true:*

- $card(A_j) \leq M_3 2^{j\gamma}$;
- *For each* $k \in A_j$, *there exist constants* $a_0 = f(2^{-j}k), a_1, ..., a_{D-1}$ *such that for all* $x \in [2^{-j}k, 2^{-j}(k+v)]$,

$$|f(x) - \sum_{m=0}^{D-1} a_m (x - 2^{-j}k)^m| \leq M_1 2^{-j\alpha_1}$$

- *For each* $k \notin A_j$, *there exist constants* $a_0 = f(2^{-j}k), a_1, ..., a_{D-1}$ *such that for all* $x \in [2^{-j}k, 2^{-j}(k+v)]$,

$$|f(x) - \sum_{m=0}^{D-1} a_m (x - 2^{-j}k)^m| \leq M_2 2^{-j\alpha}$$

Roughly speaking, the intervals with indices in S_j are "bad" intervals which contain less smooth parts of the function. The number of the "bad" intervals is controlled by M_3 and γ so that the irregular parts do not overwhelm the fundamental structure of the function.

Define the traditional Hölder classes $\Lambda^\alpha(M)$ as: for $\alpha \leq 1$,

$$\Lambda^\alpha(M) = \{f : |f(x) - f(y)| \leq M |x - y|^\alpha\}$$

and for $\alpha > 1$, $m = \lfloor \alpha \rfloor$ and $\alpha' = \alpha - m$,

$$\Lambda^\alpha(M) = \{f : |f^{(m)}(x) - f^{(m)}(y)| \leq M |x - y|^{\alpha'}\}$$

Remark 1: The function class $\mathcal{F}(\alpha_1, \alpha, \gamma, M_1, M, M_3, D, \nu)$ contains the Hölder class $\Lambda^\alpha(M)$ as a subset for any given $\alpha_1, \gamma, M_1, M_3, D, \nu$.

Remark 2: A function $f \in \mathcal{F}(\alpha_1, \alpha, \gamma, M_1, M_2, M_3, D, \nu)$ can be regarded as the sum of a regular smooth function f_0 in a Besov class $B_{\alpha\infty\infty}(M_2)$ and an irregular perturbation τ.

$$f = f_0 + \tau$$

The perurbation τ can be, for example, jump discontinuities or high frequency oscillations. For further details about the function class $\mathcal{F}$, the readers are referred to Hall, Kerkyacharian and Picard ([13]).

The following results on the wavelet coefficients of functions in $\mathcal{F}$ follows directly from the definition of $\mathcal{F}$.

Lemma 1 *Let $f \in \mathcal{F}(\alpha_1, \alpha, \gamma, M_1, M_2, M_3, D, \nu)$. Assume the wavelets $\{\phi, \psi\} \in W(D)$ with $supp(\phi) = supp(\psi) \subseteq [0, \nu]$. Let $n = 2^J$. Then*

$$\begin{aligned} |\xi_{Jk} - n^{-\frac{1}{2}} f(k/n)| &\leq M_1 \|\phi\|_1 n^{-(1/2+\alpha_1)} \quad \text{if } k \in A_j; \\ |\xi_{Jk} - n^{-\frac{1}{2}} f(k/n)| &\leq M_2 \|\phi\|_1 n^{-(1/2+\alpha)} \quad \text{if } k \notin A_j; \\ |\theta_{jk}| &\leq M_1 \|\psi\|_1 2^{-j(1/2+\alpha_1)} \quad \text{if } k \in A_j; \\ |\theta_{jk}| &\leq M_2 \|\psi\|_1 2^{-j(1/2+\alpha)} \quad \text{if } k \notin A_j; \end{aligned}$$

Now suppose one has a dyadically sampled function $\{f(\frac{k}{n})\}_{k=1}^n$ with $n = 2^J$. Then one can utilize a wavelet basis to get a good approximation of the whole function f. The following is an upper bound for the approximation error.

Theorem 1 *For a given sample $\{f(\frac{k}{n})\}_{k=1}^n$, let $f_n(x) = \sum_{k=1}^n n^{-1/2} f(\frac{k}{n}) \phi_{Jk}(x)$. Assume the wavelets $\{\phi, \psi\} \in W(D)$. Then*

$$\|f_n - f\|_2^2 \leq C_* n^{-\frac{2\alpha}{1+2\alpha}} \tag{4}$$

for all $f \in \mathcal{F}(\alpha_1, \alpha, \gamma, M_1, M_2, M_3, D, \nu)$, where the constant C_* depends on the wavelets ϕ, ψ and $\alpha_1, \alpha, \gamma, M_1, M_2, M_3, D, \nu$, but not on f.

According to Lemma 1 and this theorem, one may use $n^{-1/2} f(\frac{k}{n})$ as an approximation of $\xi_{Jk} = \langle f, \phi_{Jk} \rangle$ and use $f_n(x) = \sum_{k=1}^n n^{-1/2} f(\frac{k}{n}) \phi_{Jk}(x)$ as an approximation of f. The approximation error can be bounded based on the sample size and the smoothness of the function.

5 Optimality Of The BlockShrink Procedure

The *BlockShrink* utilizes information about neighboring wavelet coefficients. The block length increases slowly as the sample size increases. As a result, the amount of information available from the data to estimate the energy of the function within a block, and making a decision about keeping or omitting all the cofficients in the block, would be more than in the case of the term-by-term threshold rule. The *BlockShrink* increases the estimation accuracy of the wavelet coefficients and so it allows convergence rates to be improved.

In the section, we show that this is in fact true. We investigate the adaptivity and spatial adaptivity of the *BlockShrink* procedure to unknown degree of inhomogeneous smoothness over the function classes $\mathcal{F}(\alpha_1, \alpha, \gamma, M_1, M_2, M_3, D, \nu)$. We begin with global estimation.

5.1 Global Estimation

In global estimation, one is interested in estimating the whole function f or estimating the value of f at all the sample points $\{f(x_i), i = 1, \cdots, n\}$. Denote the minimax risk over function class $\mathcal{F}$ by

$$R(\mathcal{F}, n) = \inf_{\hat{f}_n} \sup_{\mathcal{F}} E\|\hat{f}_n - f\|^2$$

It is well known that the optimal rate of convergence for global estimation over Hölder class $\Lambda^\alpha(M)$ is $n^{-\frac{2\alpha}{1+2\alpha}}$. Because the function class $\mathcal{F}(\alpha_1, \alpha, \gamma, M_1, M_2, M_3, D, \nu)$ contains $\Lambda^\alpha(M_2)$ as a subset, the convergence rate over $\mathcal{F}(\alpha_1, \alpha, \gamma, M_1, M_2, M_3, D, \nu)$ can not be better than $n^{-\frac{2\alpha}{1+2\alpha}}$. Theorem 2 shows that the *BlockShrink* estimators attain the convergence rate of $n^{-\frac{2\alpha}{1+2\alpha}}$. Therefore, the estimators achieve the optimal rates across a whole interval of the function classes $\mathcal{F}$.

Theorem 2 *Suppose the wavelets $\{\phi, \psi\} \in W(D)$ and $supp(\phi) = supp(\psi) = (0, N)$. Let $\mathcal{F} = \mathcal{F}(\alpha_1, \alpha, \gamma, M_1, M_2, M_3, D, \nu)$. Then the BlockShrink estimators satisfy*

$$\sup_{f \in \mathcal{F}} E\|\hat{f}_n^* - f\|^2 \leq Cn^{-\frac{2\alpha}{1+2\alpha}}(1 + o(1)) \tag{5}$$

for all $0 < \alpha \leq D$ and for all $\nu \geq N$.

Thus, the *BlockShrink* estimator, without knowing the a priori degree or amount of smoothness of the underlying function, attains the optimal convergence rate that one could achieve by knowing the regularity.

$$\sup_{f \in \mathcal{F}} E\|\hat{f}_n^* - f\|^2 \asymp R(\mathcal{F}, n)$$

As a special case, the *BlockShrink* attains the exact optimal rates over a wide range of the traditional Hölder classes $\Lambda^\alpha(M)$:

Theorem 3 *Let the wavelets $\{\phi, \psi\} \in W(D)$. Then the BlockShrink estimators are simultaneously near minimax:*

$$\sup_{f \in \Lambda^\alpha(M)} E\|\hat{f}_n^* - f\|^2 \leq Cn^{-\frac{2\alpha}{1+2\alpha}}(1 + o(1)) \tag{6}$$

for all $0 < \alpha \leq D$ and all $M \in (0, \infty)$.

Similar results hold for the *BlockShrink* estimator of $\{f(x_i), i = 1, \cdots, n\}$.

Theorem 4 *Under the conditions of Theorem 2, the BlockShrink estimator $\hat{f}$ of $\{f(x_i), i = 1, \cdots, n\}$ is simultaneously rate-optimal:*

$$\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} E(\widehat{f_n(x_i)} - f(x_i))^2 \leq Cn^{-\frac{2\alpha}{1+2\alpha}}(1 + o(1)) \tag{7}$$

for all $0 < \alpha \leq D$ and for all $\nu \geq N$.

Remark 3 (*Generalization*): It is not difficult to generalize the adaptive results over similar function classes that allow different types of irregular perturbations in one function.

Remark 4 (*Use of Coiflets*): If one is willing to impose the following local Lipschitz condition on $\mathcal{F}$ when $\mathcal{F}$ is relatively smooth, then there is no need to use Coiflets.
(i). If $\alpha > 1 \geq \alpha_1$, then for $k \notin A_j$,

$$|f(x) - f(2^{-j}k)| \leq M_4 2^{-j}, \qquad \text{for } x \in [2^{-j}k, 2^{-j}(k+v)]$$

(ii). If $\alpha > \alpha_1 > 1$, then

$$|f(x) - f(2^{-j}k)| \leq M_4 2^{-j}, \qquad \text{for } x \in [2^{-j}k, 2^{-j}(k+v)]$$

5.2 Estimation At A Point

In global estimation, *BlockShrink* achieves complete success of adaptation across a range of function classes $\mathcal{F} = \mathcal{F}(\alpha_1, \alpha, \gamma, M_1, M_2, M_3, D, \nu)$ in terms of convergence rate. That is, one can do as well by using the *BlockShrink* procedure when the degree of smoothness is unknown as one could do if the degree of smoothness is known.

But for local estimation, i.e. estimation at a point, it is impossible to achieve complete adaptation for free. One must pay a price for adaptation. Denote the minimax risk for estimating functions at a point x_0 over a function class $\mathcal{F}$ by

$$R(\mathcal{F}, x_0, n) = \inf_{\hat{f}_n} \sup_{\mathcal{F}} E(\hat{f}_n(x_0) - f(x_0))^2$$

Consider the Hölder class $\Lambda^\alpha(M)$. The optimal rate of convergence for estimating $f(x_0)$ with α known is n^{-r} where

$$r = \frac{2\alpha}{1+2\alpha}$$

Brown and Low ([2]) and Lepski ([16]) showed that one has to pay a price for adaptation of at least a logarithmic factor even when α is known to be one of two values. They showed that the best one can do is $(\log n/n)^r$ when the smoothness index α is unknown. We call $(\log n/n)^r$ the adaptive minimax rate over the Hölder class $\Lambda^\alpha(M)$.

The following result shows that the *BlockShrink* achieves the adaptation with the minimal cost for estimating f at a point.

Theorem 5 *Let the wavelets $\{\phi, \psi\} \in W(D)$ with $D \geq \alpha$. Let $x_0 \in (0,1)$ be fixed. Then the BlockShrink estimator $\hat{f}_n^*(x_0)$ of $f(x_0)$ satisfies*

$$\sup_{f \in \Lambda^\alpha(M)} E(\hat{f}_n^*(x_0) - f(x_0))^2 \leq C \cdot (\frac{\log n}{n})^{\frac{2\alpha}{1+2\alpha}}(1 + o(1)) \tag{8}$$

Hence, $\hat{f}_n^*(x_0)$ achieves the adaptive minimax risk for a wide range of Hölder classes. The logarithmic penalty is a price one has to pay for not knowing the smoothness of the target function f. It cannot be further reduced according to the results of Brown and Low ([2]) and Lepski ([16]). Therefore, for local estimation, (8) is the optimal adaptive result one can expect in terms of convergence rates.

6 Simulation Results

The *BlockShrink* procedure is very easy to implement and the total computational cost of the procedure is of order $O(n)$. We implemented the *BlockShrink* procedure in the statistical software package S+Wavelets.

A simulation study ([4]) was conducted to investigate the performance of the *BlockShrink*. In [4], the *BlockShrink* is compared with the Donoho and Johnstone's VisuShrink, RiskShrink and SureShrink. The VisuShrink and the RiskShrink are term-by-term thresholding procedures, they differ only in threshold. The readers are referred to Donoho and Johnstone ([7]) for further details. The SureShrink thresholds the empirical wavelet coefficients by minimizing the Stein's unbiased estimate of risk at each resolution level (see [9]).

We studied eight functions representing different level of spatial variability. For each of the eight objects under study, Four different methods, the *BlockShrink*, the VisuShrink, the RiskShrink and the SureShrink, were applied to noisy versions of the data. Sample sizes from $n = 512$ to $n = 8192$ and signal-to-noise ratio $SNR = 4$ and $SNR = 7$ were studied. And several different wavelets were used.

The *BlockShrink* consistently outperforms the VisuShrink in all examples. In many cases, *BlockShrink* has better precisions with sample size n than the VisuShrink with sample size $2 \cdot n$ for all n from 512 to 8192. Furthermore, when the signal-to-noise ratio is high (e.g. $SNR = 7$), the *BlockShrink* yields uniormly better results than each of the VisuShrink, the RiskShrink and the SureShrink. The readers are refered to [4] for further details.

For the reasons of space, we report in Table 1 some of the simulation results on the Donoho and Johnstone's four test functions: Doppler, HeaviSine, Bumps and Blocks. Table 1 reports the average squared error over 100 replications with sample sizes ranging from $n = 512$ to $n = 8192$. The SNR is 7 and the wavelet is *Symmlet* 8. Figures 3 to 6 compare the visual quality of the four different reconstruction methods. All the figures were produced with the sample size 1024, the wavelet *Symmlet* 8 and the signal-to-noise-ratio 7.

The *BlockShrink* estimators are visually appealing. The reconstruction is smooth where the underlying function is smooth. They do not contain spurious fine-scale structure that are often contained in the RiskShrink estimators and the SureShrink estimators. The *BlockShrink* adapts very well to the subtle changes of the target functions. For instance, one can see from the reconstructions of Doppler and Bumps, the *BlockShrink* estimators reach to the peaks deeper than the VisuShrink estimators.

Finally, we emphasize that the same S+Wavelets program, with the same parameters, produce all four reconstructions using the four different methods; no user intervention was permitted or required.

7 Proofs

7.1 Proof of Theorem 2:

Let

$$\tilde{\Theta} = (\tilde{\xi}_{j_0 1}, \cdots, \tilde{\xi}_{j_0 2^{j_0}}, \tilde{\theta}_{j_0 1}, \cdots, \tilde{\theta}_{j_0 2^{j_0}}, \cdots, \tilde{\theta}_{J-1,1}, \cdots, \tilde{\theta}_{J-1,2^{J-1}})^T$$

be the discrete wavelet transform of $\{n^{-\frac{1}{2}} y_i\}$. Let $\tilde{f}(x) = \sum_{i=1}^{n} n^{-\frac{1}{2}} y_i \phi_{Ji}(x)$. Then

$$\tilde{f}(x) = \sum_{k=1}^{2^{j_0}} \tilde{\xi}_{j_0 k} \phi_{j_0 k}(x) + \sum_{j=j_0}^{J-1} \sum_{k=1}^{2^j} \tilde{\theta}_{jk} \psi_{jk}(x)$$

We can also write $\tilde{f}(x)$ as

$$\begin{aligned}
\tilde{f}(x) &= \sum_{i=1}^{n} [n^{-\frac{1}{2}} f(x_i) + n^{-\frac{1}{2}} \epsilon z_i] \phi_{Ji}(x) \\
&= \sum_{i=1}^{n} [\xi_{Ji} + (n^{-\frac{1}{2}} f(x_i) - \xi_{Ji}) + n^{-\frac{1}{2}} \epsilon z_i] \phi_{Ji}(x) \\
&= \sum_{k=1}^{2^{j_0}} [\xi_{j_0 k} + \tilde{a}_{j_0 k} + n^{-\frac{1}{2}} \epsilon \tilde{z}_{j_0 k}] \phi_{j_0 k}(x) + \sum_{j=j_0}^{J-1} \sum_{k=1}^{2^j} [\theta_{jk} + a_{jk} + n^{-\frac{1}{2}} \epsilon z_{jk}] \psi_{jk}(x)
\end{aligned}$$

Here, $\xi_{j_0 k}$ and θ_{jk} are the orthogonal transform of $\{\xi_{Ji}\}$ via W, and likewise $\tilde{a}_{j_0 k}$ and a_{jk} the transform of $\{n^{-\frac{1}{2}} f(x_i) - \xi_{Ji}\}$, and $\tilde{z}_{j_0 k}$ and z_{jk} the transform of $\{z_i\}$.

We note that $\tilde{z}_{j_0 k}$ and z_{jk} are i.i.d. N(0, 1) and

$$\begin{aligned}
\tilde{\xi}_{j_0 k} &= \xi_{j_0 k} + \tilde{a}_{j_0 k} + n^{-\frac{1}{2}} \epsilon \tilde{z}_{j_0 k} \\
\tilde{\theta}_{jk} &= \theta_{jk} + a_{jk} + n^{-\frac{1}{2}} \epsilon z_{jk}
\end{aligned}$$

It follows from Lemma 1 and the orthogonality of the discrete wavelet transform that

$$\sum_{k=1}^{2^{j_0}} \tilde{a}_{j_0 k}^2 + \sum_{j=j_0}^{J-1} \sum_{k=1}^{2^j} a_{jk}^2 = \sum_{i=1}^{n} (n^{-\frac{1}{2}} f(x_i) - \xi_{Ji})^2 \leq C' n^{-\frac{2\alpha}{1+2\alpha}} \tag{9}$$

Let

$$\hat{\xi}_{j_0 k} = \tilde{\xi}_{j_0 k}$$

Group the noisy wavelet coefficients into blocks of length $L = [\log n]$, the integer part of $\log n$. For simplicity, we use L $= \log n$ in the proof. Then the estimate of θ_{jk} in block (jb) is

$$\hat{\theta}_{jk} = \tilde{\theta}_{jk} I(\tilde{B}_{jb} > 5Ln^{-1}\epsilon^2)$$

The *BlockShrink* estimator of f is

$$\hat{f}_n^*(x) = \sum_k \hat{\xi}_{j_0 k} \phi_{j_0 k}(x) + \sum_{j=j_0}^{J-1} \sum_k \hat{\theta}_{jk} \psi_{jk}(x)$$

By the isometry of the function norm and the sequence norm, the risk of $\hat{f}_n^*$ can be written as

$$E\|\hat{f}_n^* - f\|_2^2 = \sum_k E(\hat{\xi}_{j_0k} - \xi_{j_0k})^2 + \sum_{j=j_0}^{J-1}\sum_k E(\hat{\theta}_{jk} - \theta_{jk})^2 + \sum_{j=J}^{\infty}\sum_k \theta_{jk}^2 \tag{10}$$

It follows from Lemma 1 that

$$\sum_{j=J}^{\infty}\sum_k \theta_{jk}^2 = o(n^{-\frac{2\alpha}{1+2\alpha}}) \tag{11}$$

Let C denote a generic constant that varies from palce to place. Then one has

$$\sum_k E(\hat{\xi}_{j_0k} - \xi_{j_0k})^2 \le 2^{j_0}\epsilon^2 n^{-1} + 2\sum_k \tilde{a}_{j_0k}^2 \le Cn^{-\frac{2\alpha}{1+2\alpha}} \tag{12}$$

Now consider $\sum_{j=j_0}^{J-1}\sum_k E(\hat{\theta}_{jk} - \theta_{jk})^2$.

$$\begin{aligned}
\sum_{j=j_0}^{J-1}\sum_k E(\hat{\theta}_{jk} - \theta_{jk})^2 &= \sum_{j=j_0}^{J-1}\sum_b\sum_{k\in(jb)} E(\hat{\theta}_{jk} - \theta_{jk})^2 \\
&\le 2n^{-1}\epsilon^2 \sum_{j=j_0}^{J-1}\sum_b E(\sum_{k\in(jb)} z_{jk}^2 I(\tilde{B}_{(jb)} > 5Ln^{-1}\epsilon^2)) \\
&\quad + \sum_{j=j_0}^{J-1}\sum_b (\sum_{k\in(jb)} \theta_{jk}^2 P(\tilde{B}_{(jb)} \le 5Ln^{-1}\epsilon^2)) + 2\sum_{j=j_0}^{J-1}\sum_k a_{jk}^2
\end{aligned}$$

Denote

$$\begin{aligned}
R_{(jb)} &= E(\sum_{k\in(jb)} z_{jk}^2 I(\tilde{B}_{(jb)} > 5Ln^{-1}\epsilon^2)) \\
R'_{(jb)} &= \sum_{k\in(jb)} \theta_{jk}^2 P(\tilde{B}_{(jb)} \le 5Ln^{-1}\epsilon^2)
\end{aligned}$$

and

$$S_1 = n^{-1}\epsilon^2 \sum_{j=j_0}^{J-1}\sum_b R_{(jb)}, \qquad S_2 = \sum_{j=j_0}^{J-1}\sum_b R'_{(jb)}$$

Therefore,

$$\sum_{j=j_0}^{J-1}\sum_k E(\hat{\theta}_{jk} - \theta_{jk})^2 \le 2S_1 + 2S_2 + 2C'n^{-\frac{2\alpha}{1+2\alpha}}$$

Let

$$\begin{aligned}
G_j &= \{\text{blocks at level j contain no coefficients with indices in } A_j\} \\
G'_j &= \{\text{blocks at level j contain at least one coefficient with indices in } A_j\} \\
G'' &= \{\text{blocks (jb) such that } \textstyle\sum_{k\in(jb)} a_{jk}^2 \ge \frac{1}{320}n^{-1}\epsilon^2\}
\end{aligned}$$

We will first assume the following two lemmas. The proof of the lemmas is given at the end.

Lemma 2

$$(i). \qquad card(G'_j) \le M_3 2^{j\gamma};$$
$$(ii). \qquad card(G'') \le CL^{-1}n^{\frac{1}{1+2\alpha}};$$

Lemma 3 *(i). If* $B_{(jb)} > 20Ln^{-1}\epsilon^2$, *then*

$$P(\tilde{B}_{(jb)} \le 5Ln^{-1}\epsilon^2) \le n^{-1} \tag{13}$$

(ii). If $B_{(jb)} \le \frac{1}{80}Ln^{-1}\epsilon^2$, *then*

$$R_{(jb)} = E(\sum_{k\in(jb)} z_{jk}^2 I(\tilde{B}_{(jb)} > 5Ln^{-1}\epsilon^2)) \le 5Ln^{-1} \tag{14}$$

Now let J_0 be an integer satisfying

$$2^{J_0} = C_0 n^{\frac{1}{1+2\alpha}}$$

where the constant $C_0 \ge (320M_2^2\|\psi\|_1^2\epsilon^{-2})^{\frac{1}{1+2\alpha}}$ so that

$$\theta_{jk}^2 \le \frac{1}{320}n^{-1}\epsilon^2 \qquad \text{for } j \ge J_0 \text{ and } k \notin A_j$$

Similarly, J_1 is chosen so that

$$2^{J_1} = C_1 n^{\frac{1}{1+2\alpha_1}}$$

and for $j \ge J_1$ and all k, $\theta_{jk}^2 \le \frac{1}{320}n^{-1}\epsilon^2$

Now, decompose S_1 into four parts:

$$\begin{aligned} S_1 &= n^{-1}\epsilon^2 \sum_{j=j_0}^{J_0-1}\sum_{(jb)} R_{(jb)} + n^{-1}\epsilon^2 \sum_{j=J_0}^{J-1}\sum_{(jb)\in G''} R_{(jb)} \\ &\quad + n^{-1}\epsilon^2 \sum_{j=J_0}^{J-1}\sum_{(jb)\in G_j\setminus G''} R_{(jb)} + n^{-1}\epsilon^2 \sum_{j=J_0}^{J-1}\sum_{(jb)\in G'_j\setminus G''} R_{(jb)} \\ &\equiv S_{11} + S_{12} + S_{13} + S_{14} \\ S_{11} &= n^{-1}\epsilon^2 \sum_{j=j_0}^{J_0-1}\sum_{(jb)} R_{(jb)} \le n^{-1}\epsilon^2 \sum_{j=j_0}^{J_0-1}\sum_{(jb)} L \le Cn^{-\frac{2\alpha}{1+2\alpha}} \end{aligned} \tag{15}$$

It follows from Lemma 2 that

$$S_{12} = n^{-1}\epsilon^2 \sum_{j=J_0}^{J-1}\sum_{(jb)\in G''} R_{(jb)} \le n^{-1}\epsilon^2 \sum_{j=J_0}^{J-1}\sum_{(jb)\in G''} L \le n^{-1}\epsilon^2 CL^{-1}n^{\frac{1}{1+2\alpha}}L \le Cn^{-\frac{2\alpha}{1+2\alpha}} \tag{16}$$

Now for $j \geq J_0$ and $(jb) \in G_j \setminus G''$,

$$B_{(jb)} = \sum_{(jb)} (\theta_{jk} + a_{jk})^2 \leq 2 \sum_{(jb)} \theta_{jk}^2 + 2 \sum_{(jb)} a_{jk}^2 \leq \frac{1}{80} L n^{-1} \epsilon^2$$

So, it follows from Lemma 3 that

$$S_{13} = n^{-1}\epsilon^2 \sum_{j=J_0}^{J-1} \sum_{(jb) \in G_j \setminus G''} R_{(jb)} \leq n^{-1}\epsilon^2 \sum_{j=J_0}^{J-1} \sum_{(jb) \in G_j \setminus G''} 5Ln^{-1} \leq Cn^{-1} \tag{17}$$

We can further decompose S_{14} into two terms:

$$S_{14} = n^{-1}\epsilon^2 \sum_{j=J_0}^{J_1-1} \sum_{(jb) \in G'_j \setminus G''} R_{(jb)} + n^{-1}\epsilon^2 \sum_{j=J_1}^{J-1} \sum_{(jb) \in G'_j \setminus G''} R_{(jb)}$$

From Lemma 2, $card(G'_j) \leq M_3 2^{j\gamma}$, so

$$n^{-1}\epsilon^2 \sum_{j=J_0}^{J_1-1} \sum_{(jb) \in G'_j \setminus G''} R_{(jb)} \leq n^{-1}\epsilon^2 \sum_{j=J_0}^{J_1-1} \sum_{(jb) \in G'_j} L \leq n^{-1}\epsilon^2 \sum_{j=J_0}^{J_1-1} M_3 2^{j\gamma} L = o(n^{-\frac{2\alpha}{1+2\alpha}}) \tag{18}$$

And for $j \geq J_1$ and $(jb) \in G_j \setminus G''$, again we have

$$B_{(jb)} = \sum_{(jb)} (\theta_{jk} + a_{jk})^2 \leq 2 \sum_{(jb)} \theta_{jk}^2 + 2 \sum_{(jb)} a_{jk}^2 \leq \frac{1}{80} L n^{-1} \epsilon^2$$

So,

$$n^{-1}\epsilon^2 \sum_{j=J_1}^{J-1} \sum_{(jb) \in G'_j \setminus G''} R_{(jb)} = o(n^{-\frac{2\alpha}{1+2\alpha}}) \tag{19}$$

Now it follows from (18) and (19) that

$$S_{14} = o(n^{-\frac{2\alpha}{1+2\alpha}}) \tag{20}$$

By putting (15) – (20) together, we have

$$S_1 \leq Cn^{-\frac{2\alpha}{1+2\alpha}}(1 + o(1)) \tag{21}$$

Now let us consider S_2. We separate S_2 into four parts:

$$\begin{aligned} S_2 &= \sum_{j=j_0}^{J-1} \sum_{(jb) \in G_j} R'_{(jb)} + \sum_{j=j_0}^{J-1} \sum_{(jb) \in G'_j} R'_{(jb)} \\ &= \Big(\sum_{j=j_0}^{J_0-1} \sum_{(jb) \in G_j} R'_{(jb)} + \sum_{j=J_0}^{J-1} \sum_{(jb) \in G_j} R'_{(jb)}\Big) + \Big(\sum_{j=j_0}^{J_1-1} \sum_{(jb) \in G'_j} R'_{(jb)} + \sum_{j=J_1}^{J-1} \sum_{(jb) \in G'_j} R'_{(jb)}\Big) \\ &\equiv S_{21} + S_{22} + S_{23} + S_{24} \end{aligned}$$

Apply Lemma 3, we have

$$\begin{aligned}
S_{21} &= \sum_{j=j_0}^{J_0-1}\sum_{(jb)\in G_j} R'_{(jb)} I(B_{(jb)} > 20Ln^{-1}\epsilon^2) + \sum_{j=j_0}^{J_0-1}\sum_{(jb)\in G_j} R'_{(jb)} I(B_{(jb)} \le 20Ln^{-1}\epsilon^2) \\
&\le \sum_{j=j_0}^{J_0-1}\sum_{(jb)\in G_j}\sum_{k\in(jb)} \theta_{jk}^2 n^{-1} + \sum_{j=j_0}^{J_0-1}\sum_{(jb)\in G_j} (40Ln^{-1}\epsilon^2 + 2\sum_{k\in(jb)} a_{jk}^2) \\
&\le Cn^{-\frac{2\alpha}{1+2\alpha}}(1+o(1)) \qquad (22) \\
S_{22} &= \sum_{j=J_0}^{J-1}\sum_{(jb)\in G_j} R'_{(jb)} \le \sum_{j=J_0}^{J-1}\sum_{(jb)\in G_j}\sum_{k\in(jb)} \theta_{jk}^2 \le Cn^{-\frac{2\alpha}{1+2\alpha}} \qquad (23) \\
S_{23} &= \sum_{j=j_0}^{J_1-1}\sum_{(jb)\in G'_j} R'_{(jb)} I(B_{(jb)} > 20Ln^{-1}\epsilon^2) + \sum_{j=j_0}^{J_1-1}\sum_{(jb)\in G'_j} R'_{(jb)} I(B_{(jb)} \le 20Ln^{-1}\epsilon^2) \\
&\le \sum_{j=j_0}^{J_1-1}\sum_{k=1}^{2^j} \theta_{jk}^2 n^{-1} + \sum_{j=j_0}^{J_1-1}\sum_{(jb)\in G'_j} (40Ln^{-1}\epsilon^2 + 2\sum_{k\in(jb)} a_{jk}^2) \\
&\le Cn^{-1} + \sum_{j=j_0}^{J_1-1} 40Ln^{-1}\epsilon^2 M_3 2^{j\gamma} + Cn^{-\frac{2\alpha}{1+2\alpha}} \\
&\le Cn^{-\frac{2\alpha}{1+2\alpha}}(1+o(1)) \qquad (24) \\
S_{24} &= \sum_{j=J_1}^{J-1}\sum_{(jb)\in G'_j} R'_{(jb)} \le \sum_{j=J_1}^{J-1}\sum_{b\in G'_j}\sum_{k\in(jb)} \theta_{jk}^2 \le C\sum_{j=J_1}^{J-1} 2^{j\gamma}2^{-j(1+2\alpha_1)}L = o(n^{-\frac{2\alpha}{1+2\alpha}}) \qquad (25)
\end{aligned}$$

It follows from (22) through (22) that

$$S_2 \le Cn^{-\frac{2\alpha}{1+2\alpha}}(1+o(1)) \qquad (26)$$

By putting (11), (12), (21) and (26) together, we prove the theorem

$$E\|\hat{f}_n^* - f\|_2^2 \le Cn^{-\frac{2\alpha}{1+2\alpha}}(1+o(1)) \qquad (27)$$

∎

Now we shall prove Lemmas 2 and 3. First Lemma 2.

Proof of Lemma 2: (i) follows from the condition $card(A_j) \le M_3 2^{j\gamma}$. To prove (ii), note that from Lemma 1, we have

$$\sum_{j=j_0}^{J-1}\sum_{k} a_{jk}^2 \le Cn^{-\frac{2\alpha}{1+2\alpha}}$$

Let $m = card(G'')$. Then (ii) follows from

$$C_0 n^{-\frac{2\alpha}{1+2\alpha}} \ge \sum_{j=j_0}^{J-1}\sum_{k} a_{jk}^2 \ge m\frac{1}{320}Ln^{-1}\epsilon^2.$$

Before we prove Lemma 3, we note the following inequalities involving χ^2 distributions:

Lemma 4 *If* $Y \sim \chi^2_L$. *Then for* $t \geq 0$,

$$(i). \quad P(Y \geq L(1+t)) \leq [(1+t)^{-1}e^t]^{-\frac{L}{2}};$$
$$(ii). \quad E(Y \cdot I(Y \geq L(1+t))) \leq L(1+t)[(1+t)^{-1}e^t]^{-\frac{L}{2}};$$

Proof of Lemma 3:

(i). It follows from the triangle inequality $\|u+v\| \geq \|u\| - \|v\|$ that if $\sum u_i^2 \geq a^2 s$ with $a > 1$, then

$$\{\sum (u_i + v_i)^2 \leq s\} \subseteq \{\sum v_i^2 \geq (a-1)^2 s\}$$

If $B_{(jb)} > 20Ln^{-1}\epsilon^2$, then by choosing $s = 5L$ and $a = 2$, one has

$$\{\tilde{B}_{(jb)} \leq 5Ln^{-1}\epsilon^2\} = \{\sum_{k\in(jb)} (n^{1/2}\epsilon^{-1}(\theta_{jk} + a_{jk}) + z_{jk})^2 \geq 5L\} \subseteq \{\sum_{k\in(jb)} z_{jk}^2 \geq 5L\}$$

So it follows from Lemma 4 that

$$P(\tilde{B}_{(jb)} \leq 5Ln^{-1}\epsilon^2) \leq P(\sum_{k\in(jb)} z_{jk}^2 \geq 5L) \leq (\frac{e^4}{5})^{-\frac{L}{2}} \leq n^{-1}$$

(ii). Similarly, the triangle inequality $\|u+v\| \leq \|u\| + \|v\|$ implies that if $\sum u_i^2 \leq a^2 s$ with $a < 1$, then

$$\{\sum (u_i + v_i)^2 > s\} \subseteq \{\sum v_i^2 \geq (1-a)^2 s\}$$

So, if $B_{(jb)} \leq \frac{1}{80}Ln^{-1}\epsilon^2$, then by choosing $s = 5L$ and $a = \frac{1}{20}$

$$\{\tilde{B}_{(jb)} > 5Ln^{-1}\epsilon^2\} \subseteq \{\sum_{k\in(jb)} z_{jk}^2 \geq (1-\frac{1}{20})^2 5L\} \subseteq \{\sum_{k\in(jb)} z_{jk}^2 \geq 4.5125L\}$$

Hence, it follows from Lemma 4 that

$$R_{(jb)} \leq E(\sum_{k\in(jb)} z_{jk}^2 I(\sum_{k\in(jb)} z_{jk}^2 \geq 4.5125L)) \leq L4.5125(\frac{e^{3.5125}}{4.5125})^{-\frac{L}{2}} \leq 5Ln^{-1}$$

∎

Before we prove Theorem 5, let us note the following two lemmas on the bound for the wavelet coefficients of functions in a Hölder class $\Lambda^\alpha(M)$ and on the bound for the approximation error a_{jk}.

Lemma 5 *Let the wavelets* $\{\phi, \psi\} \in W(D)$, *Then for all functions* $f \in \Lambda^\alpha(M)$, *the wavelet cofficients of* f *satisfies*

$$|\theta_{jk}| \leq C' \cdot 2^{-j(\frac{1}{2}+\alpha)}$$

where the constant C' *depends on the wavelets,* α *and* M *only.*

Lemma 6 *Let* $\{a_{jk}\}$ *be the Discrete Wavelet Transform of* $\{n^{-\frac{1}{2}}f(x_i) - \xi_{Ji}\}$, *then for all* $f \in \Lambda^\alpha(M)$

$$|a_{jk}| \leq C'' n^{-\alpha} 2^{-j/2}$$

where the constant C'' *depends on the wavelets,* α *and* M, *but not on* f.

7.2 Proof of Theorem 5:

First note the following lemma.

Lemma 7 *Let X_i be random variables, then*

$$E(\sum X_i)^2 \le (\sum \sqrt{EX_i^2})^2$$

It follows from Lemma 7 that

$$\begin{aligned}
&E(\hat{f}_n^*(x_0) - f(x_0))^2 \\
&\le (\sum_{k=1}^{2^{j_0}} \sqrt{E(\hat{\xi}_{j_0k} - \xi_{j_0k})^2 \phi_{j_0k}^2(x_0)} + \sum_{j=j_0}^{J-1} \sum_{k=1}^{2^j} \sqrt{E(\hat{\theta}_{jk} - \theta_{jk})^2 |\psi_{jk}^2(x_0)|} + \sum_{j=J}^{\infty} \sum_{k=1}^{2^j} |\theta_{jk}||\psi_{jk}(x_0)|)^2 \\
&\equiv (Q_1 + Q_2 + Q_3)^2
\end{aligned}$$

Let consider the three terms separately. First note that at each resolution level j, there are at most N basis functions ψ_{jk} such that $\psi_{jk}(x_0) \neq 0$, where N is the length of the support. Therefore,

$$Q_1 = \sum_{k=1}^{2^{j_0}} \sqrt{E(\hat{\xi}_{j_0k} - \xi_{j_0k})^2 \phi_{j_0k}^2(x_0)} \le 2^{j_0/2} \|\phi\|_\infty N n^{-1/2} \epsilon \tag{28}$$

For the third term, it follows from Lemma 5 that

$$Q_3 = \sum_{j=J}^{\infty} \sum_{k=1}^{2^j} |\theta_{jk}||\psi_{jk}(x_0)| \le \sum_{j=J}^{\infty} N \|\psi\|_\infty 2^{j/2} C' 2^{-j(\frac{1}{2}+\alpha)} \le C n^{-\alpha} \tag{29}$$

Apply the inequality $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ to the second term:

$$\begin{aligned}
Q_2 &= \sqrt{E(\hat{\theta}_{jk} - \theta_{jk})^2 |\psi_{jk}^2(x_0)|} \\
&= \sum_{j=j_0}^{J-1} \sum_{k=1}^{2^j} |\psi_{jk}(x_0)| \sqrt{E(\tilde{\theta}_{jk} - \theta_{jk})^2 I(\tilde{B}_{(jb)} > 5Ln^{-1}\epsilon^2) + \theta_{jk}^2 P(\tilde{B}_{(jb)} \le 5Ln^{-1}\epsilon^2)} \\
&\le \sum_{j=j_0}^{J-1} \sum_{k=1}^{2^j} |\psi_{jk}(x_0)| \sqrt{E(\tilde{\theta}_{jk} - \theta_{jk})^2 I(\tilde{B}_{(jb)} > 5Ln^{-1}\epsilon^2)} \\
&\quad + \sum_{j=j_0}^{J-1} \sum_{k=1}^{2^j} |\psi_{jk}(x_0)||\theta_{jk}| \sqrt{P(\tilde{B}_{(jb)} \le 5Ln^{-1}\epsilon^2)} \\
&\equiv Q_{21} + Q_{22}
\end{aligned}$$

Similar as in the proof of Theorem 2, let J_0' be an integer satisfying $2^{J_0'} = C_0' n^{\frac{1}{1+2\alpha}}$, where the constant C_0' is chosen so that $(\theta_{jk} + a_{jk})^2 \le \frac{1}{80} n^{-1}\epsilon^2$, for $j \ge J_0'$. (The existence of such constant C_0' follows from Lemma 5 and Lemma 6.) Therefore, by Lemma 3,

$$E(z_{jk}^2 I(\tilde{B}_{(jb)} > 5Ln^{-1}\epsilon^2)) \le 5Ln^{-1}, \qquad \text{for } j \ge J_0'$$

Also let J_1' be an integer satisfying $2^{J_1'} \simeq (\frac{n}{\log n})^{\frac{1}{1+2\alpha}}$. Then use (9) and apply the inequality $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ to Q_{21}, one has

$$\begin{aligned}
Q_{21} &\leq \sqrt{2}n^{-1/2}\epsilon \sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j} |\psi_{jk}(x_0)|\sqrt{E(z_{jk}^2 I(\tilde{B}_{(jb)} > 5Ln^{-1}\epsilon^2))} + \sqrt{2}\sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j}|\psi_{jk}(x_0)||a_{jk}| \\
&\leq \sqrt{2}n^{-1/2}\epsilon \sum_{j=j_0}^{J_0'-1}\sum_{k=1}^{2^j} |\psi_{jk}(x_0)|\sqrt{Ez_{jk}^2} + \sqrt{2}n^{-1/2}\epsilon \sum_{j=J_0}^{J-1}\sum_{k=1}^{2^j}|\psi_{jk}(x_0)|\sqrt{5Ln^{-1}} \\
&\quad +\sqrt{2}\sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j}|\psi_{jk}(x_0)|C''n^{-\alpha}2^{-j/2} \\
&= Cn^{-\frac{\alpha}{1+2\alpha}}(1+o(1))
\end{aligned}$$

Apply Lemma 3 and Lemma 6 to Q_{22},

$$\begin{aligned}
Q_{22} &\leq \sum_{j=j_0}^{J_1'-1}\sum_{k=1}^{2^j}|\psi_{jk}(x_0)||\theta_{jk}|\sqrt{P(\tilde{B}_{(jb)} \leq 5Ln^{-1}\epsilon^2)}I(B_{(jb)} > 20Ln^{-1}\epsilon^2) \\
&\quad + \sum_{j=j_0}^{J_1'-1}\sum_{k=1}^{2^j}|\psi_{jk}(x_0)||\theta_{jk}|\sqrt{P(\tilde{B}_{(jb)} \leq 5Ln^{-1}\epsilon^2)}I(B_{(jb)} \leq 20Ln^{-1}\epsilon^2) \\
&\quad + \sum_{J_1'}^{J-1}\sum_{k=1}^{2^j}|\psi_{jk}(x_0)||\theta_{jk}| \\
&\leq \sum_{j=j_0}^{J_1'-1} N2^{j/2}\|\psi\|_\infty C' \cdot 2^{-j(\frac{1}{2}+\alpha)}n^{-1/2} + \sum_{j=j_0}^{J_1'-1} N2^{j/2}\|\psi\|_\infty(\sqrt{20n^{-1}\epsilon^2} + |a_{jk}|) \\
&\quad + \sum_{J_1'}^{J-1} N2^{j/2}\|\psi\|_\infty C' \cdot 2^{-j(\frac{1}{2}+\alpha)} \\
&= C(\frac{\log n}{n})^{\frac{\alpha}{1+2\alpha}}(1+o(1))
\end{aligned}$$

So it follows that

$$Q_2 = C(\frac{\log n}{n})^{\frac{\alpha}{1+2\alpha}}(1+o(1)) \tag{30}$$

By putting (28), (29) and (30) together, we finish the proof.

$$E(\hat{f}_n^*(x_0) - f(x_0))^2 = C(\frac{\log n}{n})^{\frac{2\alpha}{1+2\alpha}}(1+o(1)) \tag{31}$$

∎

Acknowledgements: It is a pleasure to acknowledge helpful comments by Mary Ellen Bock.

References

[1] Brown, L.D. & Low, M.G. (1992). Asymptotic Equivalence of Nonparametric Regression and White Noise. Manuscript.

[2] Brown, L.D. & Low, M.G. (1992). A Constrained Risk Inequality with Applications to Nonparametric Functional Estimations. Manuscript.

[3] Cai, T. (1996). Nonparametric Function Estimation via Wavelets. Ph.D Thesis, Cornell University.

[4] Cai, T. (1996). A Simulation Study on the Performance of The BlockShrink Estimators. Manuscript.

[5] Chui, C.K. (1992). *An Introduction to Wavelets.* Academic Press, Boston, MA.

[6] Daubechies , I. (1992). *Ten Lectures on Wavelets* SIAM: Philadelphia.

[7] Donoho, D.L. & Johnstone, I.M. (1995). Ideal Spatial Adaptation via Wavelet Shrinkage. *Biometrika,* ***81***, 425–455.

[8] Donoho, D.L. & Johnstone, I.M. (1992). Neo-Classic Minimax Problems, Tresholding, and Adaptation. Technical Report, Stanford University.

[9] Donoho, D.L. & Johnstone, I.M. (1994). Adapting to Unknown Smoothness via Wavelet Shrinkage. Technical Report, Stanford University.

[10] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. & Picard, D. (1995). Wavelet Shrinkage: Asymptopia?, *J. Roy. Stat. Soc. Ser. B,* ***57***, 301–369.

[11] Efroimovich, S.Y. (1985). Nonparametric Estimation of a Density of Unknown Smoothness. Theor. Probab. Appl. 30, 557-661.

[12] Hall, P., Kerkyacharian, G. & Picard, D. (1995a). On Block Thresholding Rules For Curve Estimation Using Wavelet Methods, manuscript.

[13] Hall, P., Kerkyacharian, G. & Picard, D. (1995b). On The Minimax Optimality of Block Thresholded Wavelet Estimators, manuscript.

[14] Hall, P., Penev, S., Kerkyacharian, G. & Picard, D. (1996). Numerical Performance of Block Thresholded Wavelet Estimators, manuscript.

[15] Kerkyacharian, G., Picard, D. & Tribouley, K (1994). L_p Adaptive Density Estimation. Technical Report, Université Paris VII.

[16] Lepski, O.V. (1990). On a Problem of Adaptive Estimation on White Gaussian Noise. *Theory of Probability and Appl.* **35**, 3, 454-466

Formulas of the Donoho and Johnstone's four test functions:

1. *Doppler.*

$$f(x) = \sqrt{x(1-x)}\sin(2.1\pi/(x+.05))$$

2. *HeaviSine.*

$$f(x) = 4\sin 4\pi x - sgn(x-.3) - sgn(.72-x)$$

3. *Bumps.*

$$f(x) = \sum h_j K((x-x_j)/w_j) \qquad K(x) = (1+|x|)^{-4}.$$

$$\begin{array}{lllllllllll}
(x_j) & = (.1, & .13, & .15, & .23, & .25, & .40, & .44, & .65, & .76, & .78, & .81) \\
(h_j) & = (4, & 5, & 3, & 4, & 5, & 4.2, & 2.1, & 4.3, & 3.1, & 5.1, & 4.2) \\
(w_j) & = (.005, & .005, & .006, & .01, & .01, & .03, & .01, & .01, & .005, & .008, & .005)
\end{array}$$

4. *Blocks.*

$$f(x) = \sum h_j K(x-x_j) \qquad K(x) = (1+sgn(x))/2.$$

$$\begin{array}{lllllllllll}
(x_j) & = (.1, & .13, & .15, & .23, & .25, & .40, & .44, & .65, & .76, & .78, & .81) \\
(h_j) & = (4, & -5, & 3, & -4, & 5, & -4.2, & 2.1, & 4.3, & -3.1, & 5.1, & -4.2)
\end{array}$$

Table 1: Mean Squared Error From 100 Replications

n	BlockShrink	VisuShrink	RiskShrink	SureShrink
Doppler				
512	0.584	1.304	0.946	1.121
1024	0.341	0.868	0.605	0.635
2048	0.195	0.560	0.368	0.379
4096	0.093	0.309	0.231	0.230
8192	0.050	0.185	0.134	0.128
HeaviSine				
512	0.369	0.598	0.588	0.596
1024	0.195	0.336	0.324	0.331
2048	0.119	0.200	0.177	0.188
4096	0.088	0.161	0.156	0.157
8192	0.045	0.086	0.081	0.082
Bumps				
512	1.187	3.512	1.895	1.240
1024	0.701	2.352	1.253	0.993
2048	0.437	1.569	0.837	0.575
4096	0.319	0.736	0.453	0.372
8192	0.179	0.454	0.279	0.220
Blocks				
512	1.118	2.158	1.289	1.381
1024	0.685	1.549	0.904	0.891
2048	0.425	1.133	0.650	0.656
4096	0.372	0.619	0.403	0.410
8192	0.211	0.440	0.276	0.285

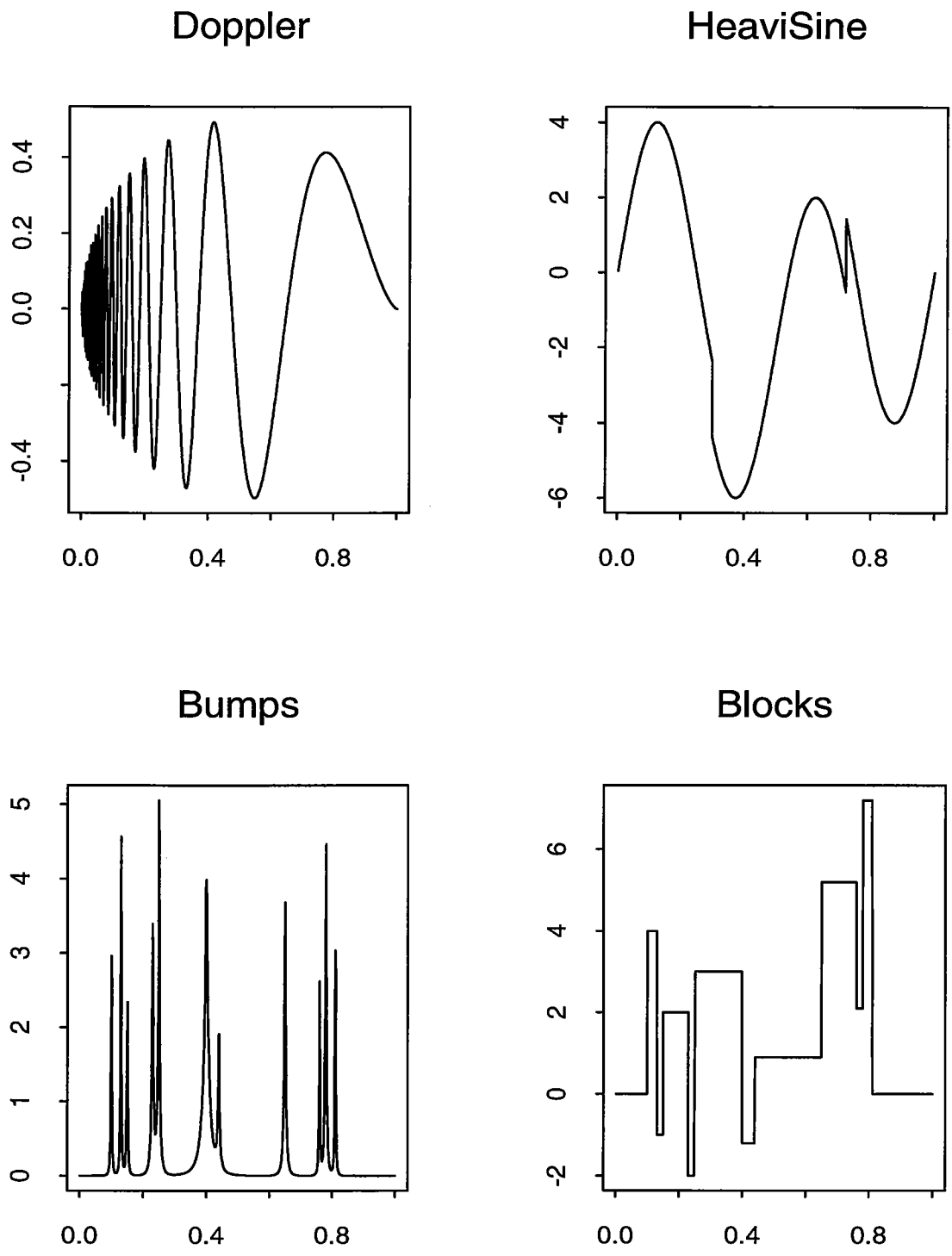


Figure 1: Test Functions

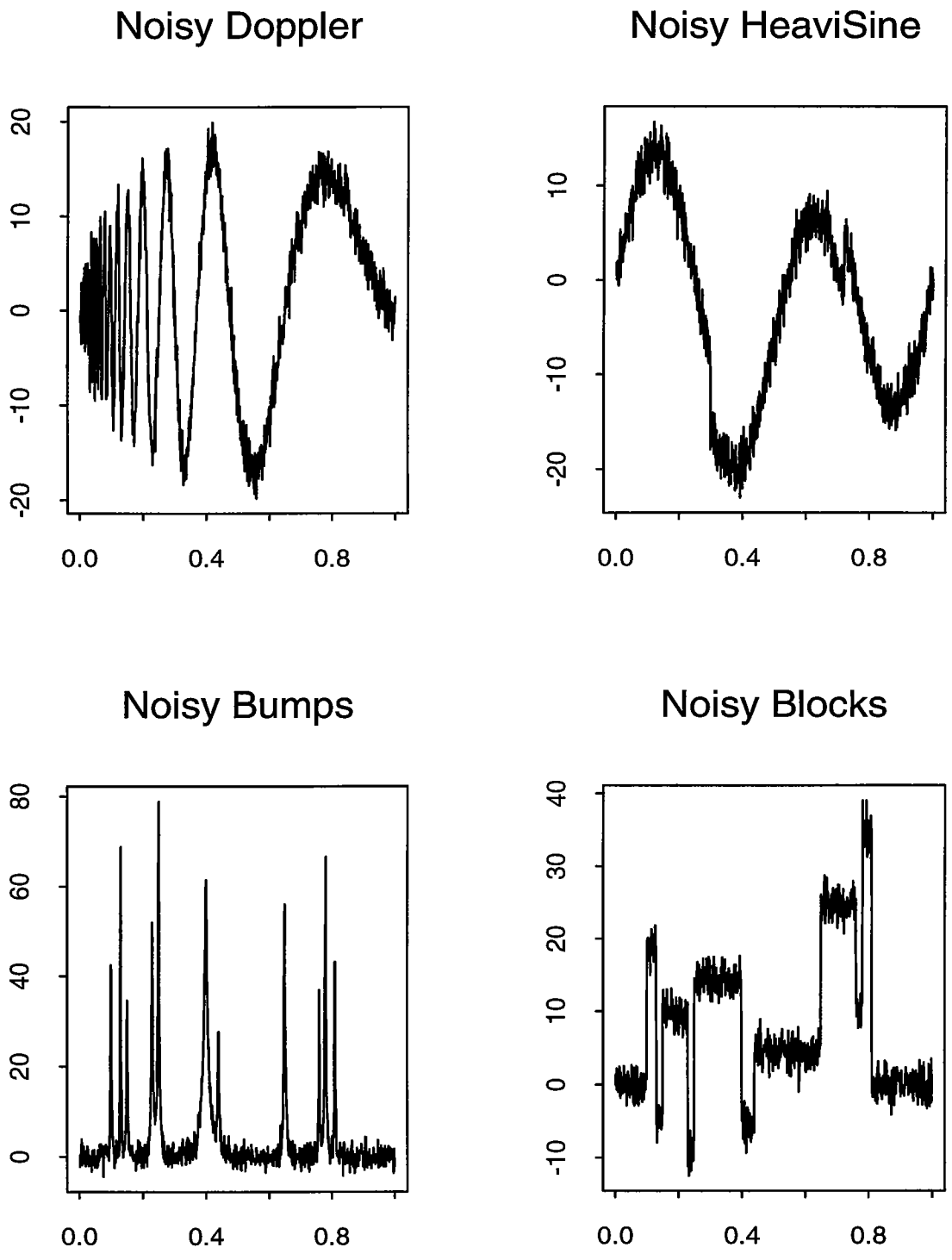


Figure 2: Test Functions + Noise

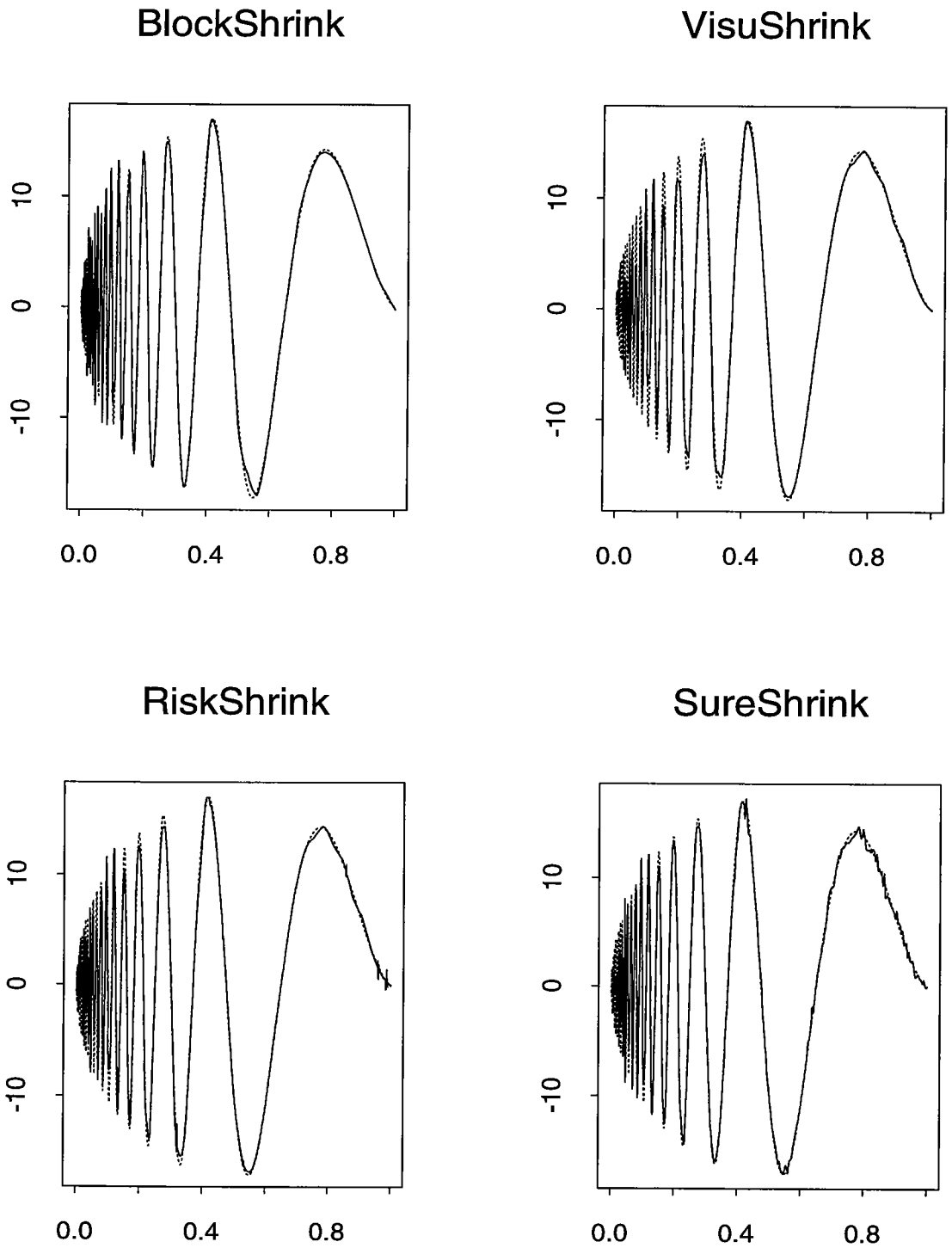


Figure 3: Reconstructions of Doppler

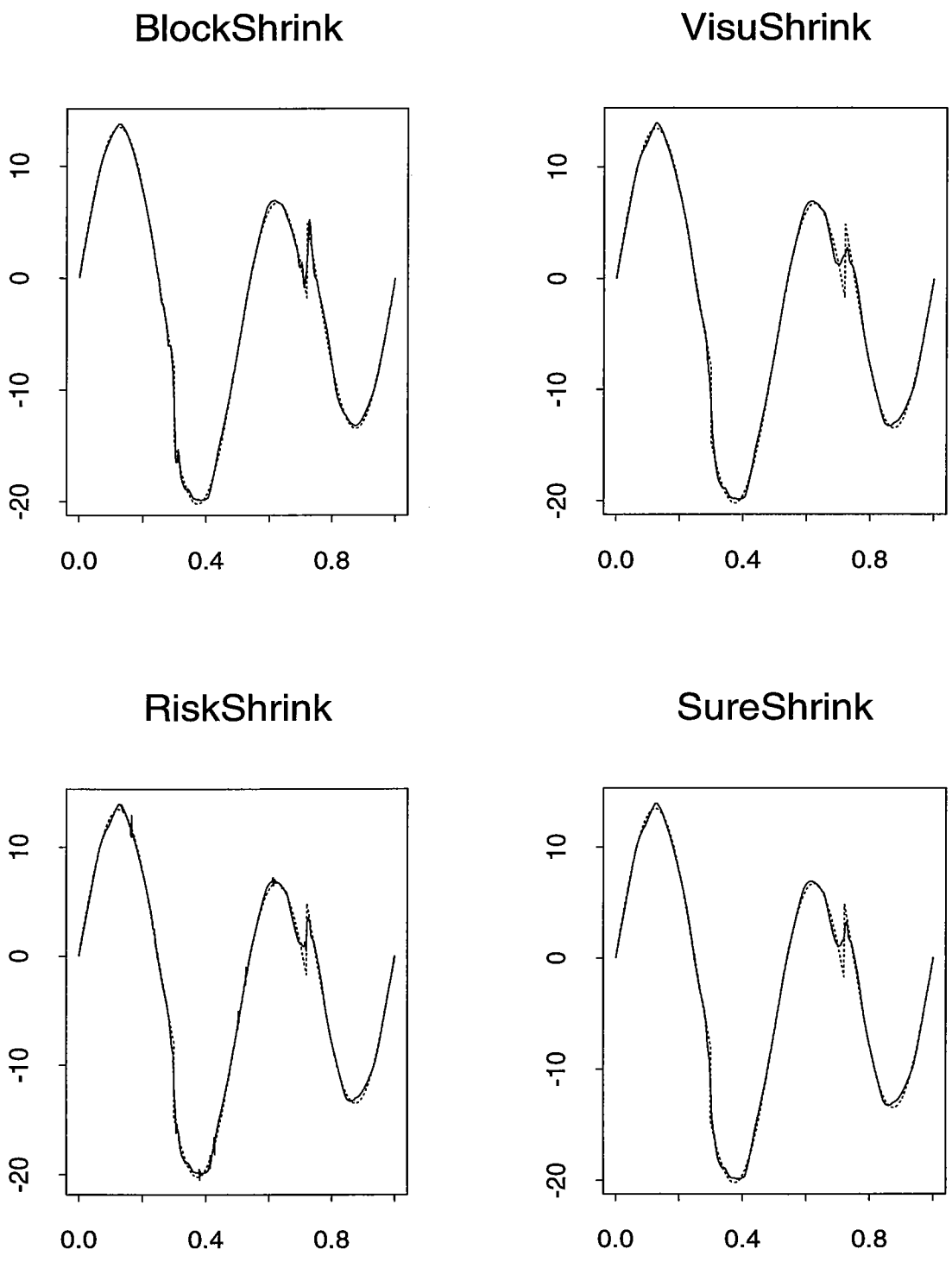


Figure 4: Reconstructions of HeaviSine

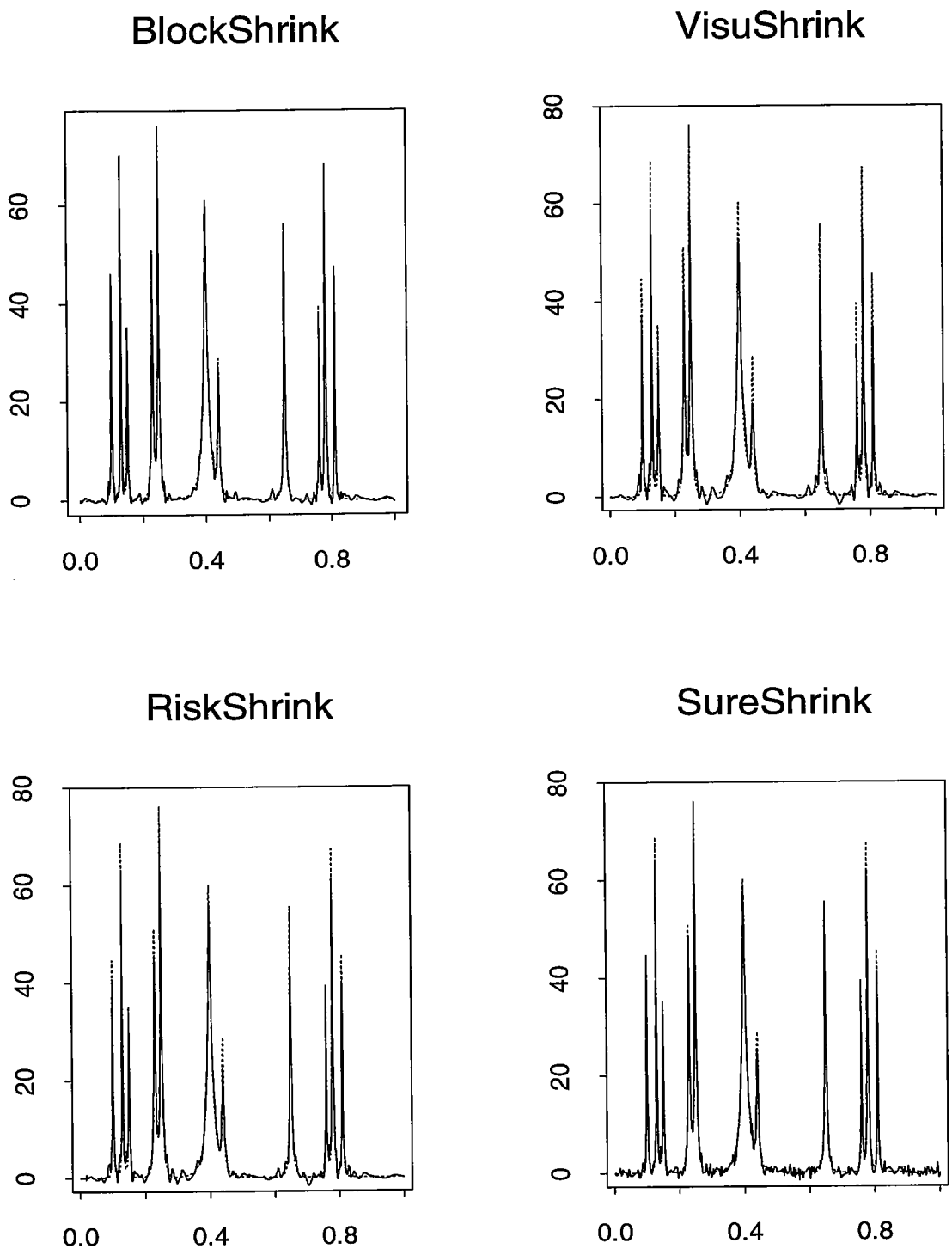


Figure 5: Reconstructions of Bumps

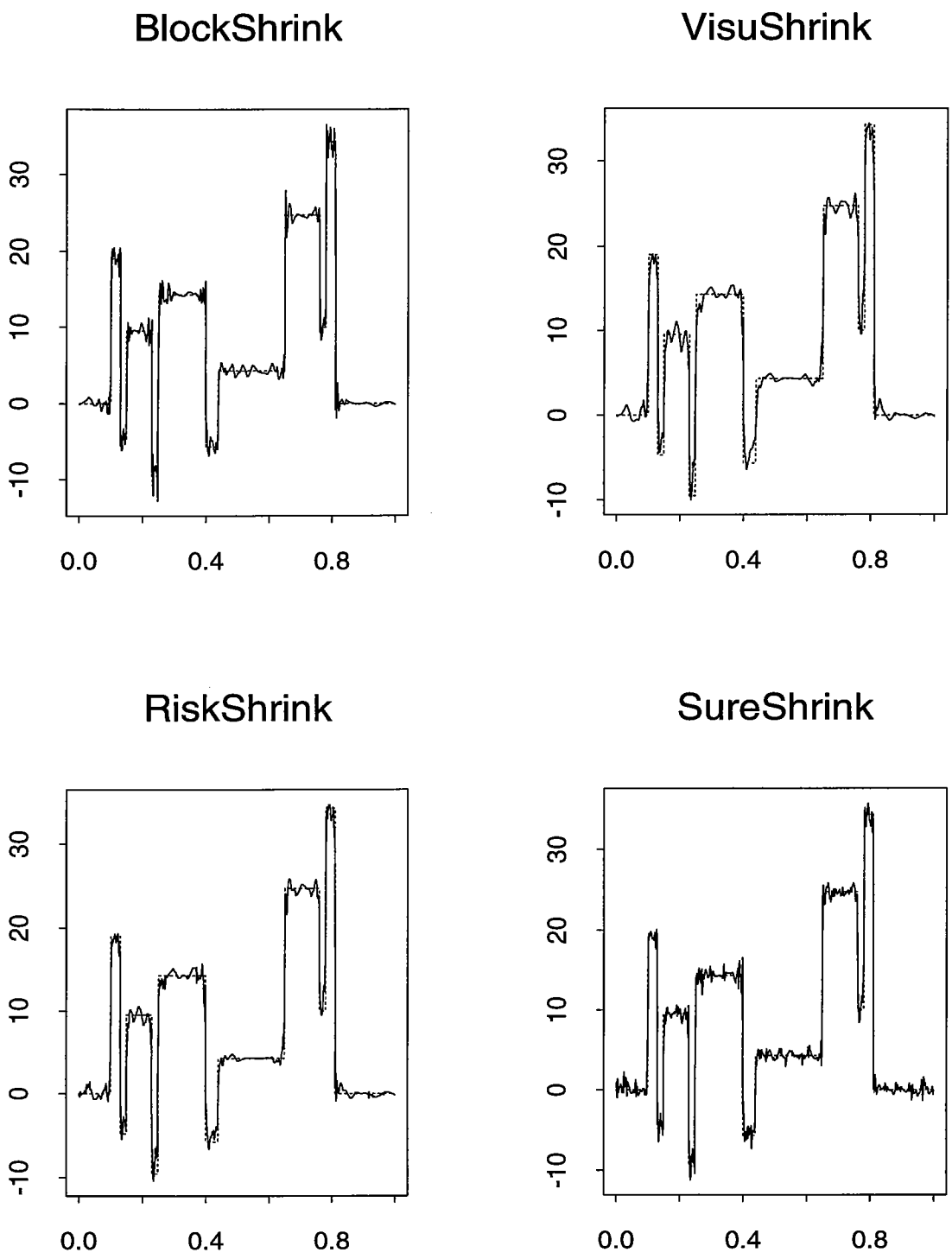


Figure 6: Reconstructions of Blocks